Performance Testing Tools: a Systematic Mapping Review

***Abstract*—The benefits of using performance testing tools have been widely discussed in software engineering. However, due to the lack of knowledge of the existing tools, these are not used or re-implemented. This systematic mapping presents the results of a study carried out with the objective of collecting and analyzing data on tools that carry out performance tests and monitoring in software aimed at computer science field. With the results of this study, it was possible to identify the amount of research related to existing performance testing tools in the literature, and also to characterize these tools. The study will also help professionals and researchers looking for new tools or tools that have certain peculiarities in their features.**

***Index Terms*—Performance testing, performance testing tools, systematic mapping study, feature model**

I. INTRODUCTION

Software testing is a key part of developing quality software products. These activities require a lot of resources, such as time and money when compared to other phases of the software development life cycle, as these can represent up to 60 % of the total development effort [1]. Software testing is laborious and very expensive, so there is a need to reduce the human-made tests and automate this process [2]. Software testing is necessary because errors are often introduced into the software unthinkingly, as they are designed and built. Software becomes increasingly complex, which means there are more lines of code and more complete tests that need to be done [3].

The purpose of testing may be quality assurance, verification, validation or reliability estimation. It is an exchange between budget, time and quality. Software quality is the central concern of software engineering. Testing is the most commonly used approach to ensuring software quality [4].

To further clarify this issue, a systematic mapping study (SMS) was carried out in order to identify and categorize existing performance testing tools in the literature and provide an overview of its characteristics, dependencies, supported scripts and technologies employed.

As discusses de pesquisa so organizadas em vrias sees. A Seo I discute x coisa. A Seo II descreve blablabla, A Seo II descreve mais alguma coisa. A Seo IV conclui a pesquisa juntamento com a taxonomia mais foda de todas.

II. BACKGROUND

lorem ipsum dolor sit amet consectetur adipiscing elit

III. METHOD

The conduction of this systematic mapping of the literature was based on the guidelines proposed on [5]. In a general context, the mapping began with the definition of the protocol, which implies in the definition of scope and objectives, the formulation of research questions, the definition of inclusion and exclusion criteria, planing the quality assessment stage on the achieved primary studies, and defining the relevant data that will be extracted from the studies that reach the final phase.

A. Scope and Objective

The objective of this study is to identify and characterize existing performance testing tools in the literature, with the purpose of assisting professionals and researchers who search for new tools or tools that have certain particularities in the development and execution of performance tests. One of the main focuses of the study is to identify the academic and open source tools and find out how they work. The description of the goal is described according to the GQM (Goal, Question, Metric) paradigm [6] and can be observed in the Table I.

TABLE I
OBJECTIVE ACCORDING TO THE GQM PARADIGM

With the purpose of:	Identify / Characterize
In relation to:	performance testing tools
From the point of view of:	performance test engineers and researchers
In the context:	of performance testing environment

B. Question Structure

The research questions (RQs) were structured based on the Population, Intervention, Comparison and Result (PICO) criteria: (a) population, published research on software; (b) intervention: performance testing; (c) comparison: not applicable; and (d) results: performance testing tools.

C. Research Questions

The following research questions were defined.

RQ1: What are the tools that support the performance testing?
RQ2: What techniques or methods of load generation do the performance testing tools implement?
RQ3: What are the structuring and organizing approaches of the test scripts interpreted by the performance load generators?
RQ4: What performance testing monitoring approaches are applied?
RQ5: What are the persistence strategies of metrics data collected by performance test monitors?

D. Search Process

Formal literature research was conducted using only databases that (i) have a search engine; (ii) has a search engine capable of using keywords; and (iii) contains computer science documents. The selection includes the bases: Association for Computing Machinery (ACM) Digital Library, Engineering Village, IEEE Xplore, ScienceDirect, SCOPUS and SpringerLink (https://www.dl.acm.org, https://www.engineeringvillage.com, https://www.ieeexplore.ieee.org, https://www.sciencedirect.com, https://www.scopus.com and https://www.link.springer.com).

To define the search string we use the terms and synonyms present in the table II, the Boolean operator 'OR' to select alternative terms and synonyms, and the Boolean operator 'AND' to add more terms to the string. The resulting string can be seen in Figure 1.

TABLE II
SEARCH STRING DEFINITION

Terms	Synonyms
Performance Test	Load Test, Stress Test, Soak Test, Spike Test, Workload Test, Automation Test
Tool	Generator, Injector, Monitor, Analyzer, Framework, Suite, Environment, Plug*in
Software	Application, System

```
TITLE("Performance Test" OR "Load Test" OR "Stress Test" OR "Spike Test" OR "Soak Test" OR "Workload Test" OR "Automation Test") AND (Tool OR Plugin OR Plug-In OR Framework OR Generator OR Monitor OR Injector OR Suite OR Analyzer OR Environment) AND (Software OR System OR Application))
```

Fig. 1. Search String

E. Inclusion and exclusion criteria

IC1: The publication should report the use of a tool that supports performance testing.
IC2: The publication should propose a tool to support performance testing.

EC1: Duplicated studies.
EC2: The publication is not related to performance testing in the software area. eg, performance test of an engine.
EC3: The publication is written in a language other than English.
EC4: The publication is only available in the form of abstract, slide show, poster or short paper.
EC5: The publication is not available for download.
EC6: The publication does not report or propose a performance testing tool.

F. Quality assessment criteria

The purpose of using quality assessment criteria is to evaluate the scientific quality of the selected studies, not measuring the literary quality, relevance or originality. The quality assessment criteria will be used in two stages, the first stage being the individual evaluation of each researcher, to reduce the probability of bias, followed by the second stage where the researchers should arrive in a consensual note about the publications that received a divergent grade.

Each of the cited QA criteria is evaluated by each researcher, according to the following degree: Yes (S) = 1; Partial (P) = 0.5; No (N) = 0 ;. So the total score totaling the five questions can result in: 0-1 (very bad); 1.5 or 2.0 (regular); 2.5 or 3.0 (good); 3.5 or 4.0 (very good); and 4.5 or 5.0 (excellent). Each of the criteria and their possible evaluations are described below:

QA1: Does the publication make a contribution to the software performance testing field?
S: A contribution is explicitly defined in the publication;
P: A contribution is implied;
N: No contributions could be identified;

QA2: Does the publication feature a software performance testing tool?
S: The publication proposes and demonstrates the use of a tool;
P: The publication proposes or demonstrates the use of a tool, never both;
N: No, the publication does not propose or demonstrate the use of a tool;

QA3: Does the publication apply any type of empirical evaluation?
S: The publication explicitly applied an evaluation (for example, a case study, an experiment or proof of correctness);
P: The evaluation is a "Toy" example;
N: No evaluations could be identified;

QA4: Does the publication present some type of analysis, showing results?
S: The publication presents some type of analysis or shows the results obtained;
P: The publication presents only a summary of the results;
N: No form of analysis or result were presented;

QA5: Does the publication describe the techniques used in load generation and monitoring?
S: The publication explicitly describes load generation and monitoring techniques.
P: The publication describes either load generation techniques, or monitoring techniques, never both.
N: The publication does not describe any load

generation or monitoring techniques.

G. Selection Process

The selection process was divided into 5 stages, which were performed by two researchers. The process steps as well as the researchers involved are described below:

1) *Initial Selection*: strings were generated using the selected keywords and synonyms. An initial selection was performed by researcher one, according to criteria EC1, EC2 and EC4 (see Section III-E);
2) *Eliminate redundancies*: at this stage, researchers one and two worked together on a pre-analysis of articles to eliminate redundancies.
3) *Intermediate selection*: at this stage, researchers one and two read separately the title and the abstract (reading the introduction and conclusion when necessary) of each study. Here, the researchers decided to select or reject an article following IC1, IC2, EC1 and EC3 (see Section III-E).
4) *Final selection and elimination of discrepancies*: At this stage, all other studies were completely read by researchers one and two, who applied the same criteria of the intermediate selection. In case of disagreement over the evaluation of the study, a third researcher would read the studies and discuss whether or not the study should be included in the final selection.
5) *Quality assessment*: Based on the quality criteria (see Section III-F), we assessed the quality of the studies that were read in the final selection stage. The quality criteria were evaluated independently by the two researchers; therefore, reducing the probability of erroneous and/or biased results, the researchers agreed in a consensual note on the publications that received a divergent grade.

H. Data Extraction

To extract the relevant data from the selected publications, we produced a form that would help to answer the RQs and also to check the QA criteria. The following data were extracted for each study:

- Title.
- Year of publication.
- Authors.
- Type of contribution: tool, approach, methodology, technique, method, or framework.
- Name of the tool / approach / methodology / technique presented.
- Type of license supported by the tool (commercial, academic, open-source).
- Language or types of script supported.
- Supported classes and types of metrics in respect to performance monitoring.
- Reports generated on the tests performed.
- Architecture and organization of data persistence.
- Domain (Web, Mobile, Desktop, etc).
- Type of performance testing: Load, Stress, Soak, Spike, Configuration, Isolation and Internet testing.

The data presented here were manipulated using the Thoth tool (which can be accessed at http://lesse.com.br/tools/slr). This tool assisted in the whole process of this review, supporting the classification and extraction of data, the selection and qualification of the papers and also aided in visualizing the results.

An important issue during data extraction was solved in a way that both researchers acted as data extractors and also as data verifiers, thus reducing the probability of errors and / or bias in data extraction.

IV. RESULTS AND DISCUSSION

In this section we discuss the answers to our RQs III-C. In each case, we indicate the utility of these results for instructors and researchers.

- RQ1: What are the tools that support the performance testing?

Foram identificadas 'x' ferramentas de teste de desempenho por meio do nosso SMS. A tabela x lista estas ferramentas. Propusemos trłs esquemas de classificao para as ferramentas: (1) por tipo de licena, (2) por tal coisa, (3) porque bonita.

V. THREATS TO VALIDITY

lorem ipsum dolor sit amet consectetur adipiscing elit

VI. PERSPECTIVES OF RESEARCH

lorem ipsum dolor sit amet consectetur adipiscing elit

VII. CONCLUSIONS

lorem ipsum dolor sit amet consectetur adipiscing elit

REFERENCES

[1] G. J. Myers and C. Sandler, *The Art of Software Testing*. USA: John Wiley & Sons, Inc., 2004.

[2] P.-L. Poon, S.-F. Tang, T. H. Tse, and T. Y. Chen, "Choc'late: A framework for specification-based testing," *Commun. ACM*, vol. 53, no. 4, pp. 113–118, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721687

[3] C. Douce, D. Livingstone, and J. Orwell, "Automatic test-based assessment of programming: A review," *J. Educ. Resour. Comput.*, vol. 5, no. 3, Sep. 2005. [Online]. Available: http://doi.acm.org/10.1145/1163405.1163409

[4] A. Jain, M. Jain, and S. Dhankar, "A comparison of ranorex and qtp automated testing tools and their impact on software testing," *IJEMS*, vol. 1, no. 1, pp. 8–12, 2014.

[5] B. A. Kitchenham, *Kitchenham, B.: Guidelines for performing Systematic Literature Reviews in software engineering. EBSE Technical Report EBSE-2007-01*, 01 2007.

[6] H. Koziolek, "Goal, question, metric," in *Dependability metrics*. Springer, 2008, pp. 39–42.